



Facial Expression Recognition from Speech

Guylaine Lejan, Nathan Souviraà-Labastie, Frédéric Bimbot

► To cite this version:

Guylaine Lejan, Nathan Souviraà-Labastie, Frédéric Bimbot. Facial Expression Recognition from Speech. [Research Report] RR-8337, INRIA. 2013. hal-00848624

HAL Id: hal-00848624

<https://inria.hal.science/hal-00848624>

Submitted on 26 Jul 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Facial expression recognition from speech

Guylaine Lejan , Nathan Souviraà-Labastie , Frédéric Bimbot

**RESEARCH
REPORT**

N° 8337

26/07/2013

Project-Team PANAMA

ISSN 0249-6399

Facial expression recognition from speech

Guylaine Lejan¹ , Nathan Souviraà-Labastie ²,

Frédéric Bimbot³

Project-Teams PANAMA

Research Report N° 8337 — 26/07/2013 —14 pages.

Abstract: In this document, we present a facial expression recognition method developed during the ReV-TV project by the Metiss team. First we detail the representation of facial expressions, then the database construction, the audio features and the classifier used.

Key-words: Facial expression recognition, speech processing

¹ INRIA Centre de Rennes – Bretagne Atlantique – guylaine.lejan@inria.fr

² IRISA/Université de Rennes 1 – nathan.souviraà-labastie@irisa.fr

³ IRISA/CNRS UMR 6074– frederic.bimbot@irisa.fr



**RESEARCH CENTRE
BRETAGNE ATALANTIQUE**

Campus universitaire de Beaulieu
35042 Rennes Cedex France

Reconnaissance d' expression faciale à partir du signal de parole

Résumé : Nous présentons dans ce document une méthode de reconnaissance d'expressions faciales développée dans le cadre du projet ReV-TV par l'équipe METISS. Nous expliquons tout d'abord pourquoi cette représentation par expressions faciales a été choisi, avant de présenter la construction de la base de données, les paramètres audio et les méthodes de classification utilisées.

Mots clés : Reconnaissance d'expressions faciales, traitement de la parole.

1. Introduction.....	7
2. Databases identification.....	7
2.1 Facial expressions definition.....	7
2.2 Databases presentation.....	7
3. Description of the method.....	8
3.1 Low level features calculated for eyebrows / neutral eyebrow database.....	8
3.2 Low level features calculated for Smile / No smile database.....	9
3.3 Low level features calculated for Head shake / No head shake database.....	9
3.4 Classification.....	9
4. Results	10
5. Discussion.....	11
5.1 Audio / video fusion.....	11
5.2 Enhancement of the audio databases.....	12
Bibliography	13

1. Introduction

Lip synchronization method developed during the ReV-TV project are able to control in real time the mouth of a virtual character (avatar) from speech but don't give information on the emotional state of the player, and thus don't allow to display emotion on the avatar's face. Emotion's identification from speech have been the subject of many publications [1] [2] but also subject of many controversy especially regarding the definition and the classification of the different emotions. Most of this publications identify the emotions in a finite number of classes and are often based on discrete emotions [3] (joy, anger, fear, sadness, surprise, disgust). These discrete emotions are sometimes difficult to recognized from players of TV-show, like anger, disgust or fear. Other studies have shown that emotions can be quantified into a continuous space instead of a discrete space [4][5][6].

To get rid of these difficulties, we present here a first prospective study based on the detection of facial characteristics of the most observed facial expressions from players of TV-show, only from speech. For this purpose, we have analyzed video sequences and identified these characteristics (shake of the head, smile, eyebrow position, ...). Then we propose a detection method.

2. Databases identification

2.1 Facial expressions definition

We decide to not recognize the 6 basic emotions (joy, anger, fear, sadness, surprise, disgust) or any other set defined with more or less subjectivity. The most frequently observed facial characteristics have been selected from the main characteristics of the face (the mouth, the eyebrows, the eyes and the head) :

- Eyebrows : 4 positions have been identified (up, frowned, lightly up, movement up to down) in case as surprise, deception or questioning. We observe frequently different positions (frowned or up) for the same emotions with different players.
- Mouth : Stretch (smile) more or less important, slant position, plout, tight-lipped, stretch toward the bottom as in the disappointment, opening more or less large.
- Shake of the head to confirm an assertion or in case of a refusal or a disagreement.
- Eyes: opening of the eyes (wide, medium, little), blinking, movement of the eyes (searching, reflecting, shy, to side)

We choose to simplify this set of facial characteristics to focus on three of them: an eyebrow position that differ from a neutral position, the smile and the shake of the head. This set of facial characteristics let us represent in a global fashion the principal emotions. Given the fact that those characteristics are not linked, we decide to create three distinct databases.

2.2 Databases presentation

We first have segmented TV-shows ("Roue de la fortune" , "Les Z' Amours" , "Tout le monde veut prendre sa place"), into sequences that contained the player's responses to the anchor. Those sequences are then classified and segmented more precisely according to the set of three main characteritics presented previously. It's worth noting that those sequences are very short especially for head shaking and eyebrows movement. Furthermore, other sequences are picked up to represent the neutral state or emotion. In the end, only the audio record is kept to create the databases. Training and test databases are extracted from different TV-show to insure a good generalization of the model.

For each frame, of 40ms with 10ms of overlap, low-level features and their first derivative are calculated. Statistical properties of frames of 500ms with 10ms overlap are then extracted. The dimension of the final feature vector is not dependent of the length of the sentence. The number of extracted frames from the different databases are given in Table 1.

Table 1. Size in term of number of samples of the used databases.

Corpus	Training set	Test set
Smile/ No smile	4738 / 6084	1764 / 770
Eyebrows/ Neutral eyebrows	2798 / 3240	1904 / 686
Head shacking/ No head shacking	565 / 2215	802 / 1342

3. Description of the method

In this part, we present the different features relevant for the detection of each facial characteristics, respectively named "Smile" , "Eyebrows" , "Head-shaking" .

3.1 Low level features calculated for eyebrows / neutral eyebrow database

Eyebrows movement often appear with a change of intonation in the speaker voice: going up during a question or going down during an assertion. There is also tags such as "eehh" (in French) that indicate indecision or reflection. For this reason, low-level features in time and frequency domain have been calculated.

Following low level features have been calculated on eyebrow / neutral eyebrow database :

- Probability of voicing
- ZCR (Zero Crossing Rate)
- Intensity
- 12 MFCCs (Mel Frequency Cepstral Coefficients)
- 8 LSP (Line Spectral Pairs)
- Pitch (F0),
- F0 envelope.

The following 15 fonctions have been used on the 25 low level features to characterise 500ms frames :

- Valeurs extrêmes :
 - Maximum,
 - Minimum,
 - Range (max-min),
 - Relative position of the maximum,
 - Relative position of the minimum,
 - Arithmetic mean.
- Moments
 - Standard deviation,

- Skewness,
 - Kurtosis.
- Percentiles
 - The 3 quartiles (q_1, q_2, q_3),
 - Inter-quartiles range ($q_2 - q_1, q_3 - q_2, q_3 - q_1$).
- Regression terms (linear approximation)
 - Gradient a (linear regression),
 - Offset b (linear regression), a et b give the equation of the linear regression line,
 - Linear error between the contour and the linear regression line,
 - Squared error between the contour and the linear regression line.

This 15 functions are used over the first 25 features and their first derivative. In the end, the feature vector dimension is 750 ($(25+25)*15$).

3.2 Low level features calculated for Smile / No smile database

Smile detection has been the subject of a lot of research. Four frequently [7][8] used type of low level features have been selected:

- Probability of voicing
- 12 MFCCs (Mel Frequency Cepstral Coefficients)
- Pitch (F_0),
- F_0 envelope.

All the statistic functions used for eyebrow / neutral eyebrow database excepting the regression terms are also used on the Smile / No smile database. This 11 functions are used over the 15 features and their first derivative. In the end, the feature vector dimension is 330 ($(15+15)*11$).

3.3 Low level features calculated for Head shake / No head shake database

Head shake is generally observed when an assertion or a protest needs to be reinforced. The intonation of the voice change then (going down or up) quite quickly. Hence, the five following features have been selected :

- Intensity
- Probability of voicing
- ZCR (Zero Crossing Rate)
- Pitch (F_0),
- F_0 envelope.

The 15 statistic functions are used over this 5 features and their first derivative. In the end, the feature vector dimension is 150 ($(5+5)*15$).

3.4 Classification

Experiments have been done to determine which classifier would fit the best among the following classifiers :

- Multi Layer Perceptron (MLP) (a basic Artificial Neural Network (ANN)) with a number of hidden neurons equal to the sum of the number of input and output of the network divided by two (*i.e.* the size of the feature vector plus 2 divided by 2).
- A MLP with its number of hidden neurons fixed at 200 in order to limit the number of free parameters and so the number of needed training samples for the training.
- The RepTree of Weka tool associated with Bagging (Bootstrap aggregating) [9]. The RepTree is a fast variant of the C4.5 algorithm [10]. The pruning is done by a cross-validation method. The target of the bagging is to tackle the instability of the classifier. For example, if a minor change of the data cause an important change of the model.

4. Results

For each experiment ("Smile" , "Eyebrows" , "Head shake"), one classifier have been trained on the train set (*cf.* Table 1). Each classifier has two output classes (presence or absence). Then the classifiers have been tested on associated test set.

Table 2, Table 3 and Table 4 show results from these experiments :

Table 2: Classification results of the head shake detection

Head shake	MLP()	MLP(200)	Bagging +RepTree
% Correct	61.66	61.94	58.07
Precision	0.68	0.69	0.46
Recall	0.61	0.62	0.72
F-score	0.62	0.62	0.56

Table 3: Classification results of the smile detection

Smile	MLP()	MLP(200)	Bagging +RepTree
% Correct	65.15	64.05	57.18
Precision	0.69	0.68	0.69
Recall	0.65	0.64	0.70
F-score	0.66	0.65	0.69

Table 4: Classification results of the eyebrow detection

Eyebrow	MLP()	MLP(200)	Bagging +RepTree
% Correct	69.74	69.98	69.85
Precision	0.66	0.66	0.70
Recall	0.70	0.70	0.98
F-score	0.66	0.66	0.82

Differences between the different classifiers are short, with higher scores for MLPs. Results are promising and provide a new simple way to represent emotions by using the prediction of facial characteristics.

5. Discussion

5.1 Audio / video fusion

Actual video analysis is based on the detection of the discrete set of emotional states defined by Ekman[3] : joy, anger, fear, sadness, surprise, disgust. Concerning the ReV-TV project, this set can be reduce to joy, sadness, surprise, and fear. Indeed, disgust and anger are not often observed during TV-show. A specific video analysis aims at recognizing those discrete emotions. The video analysis performances can be increase by the addition of audio information, typically those presented in this document : "Smile" , "Eyebrows" , "Head shake" . Smile detection can be combined with joy detection, "eyebrow" with the surprise, sadness, or even fear detection, the "head shake" can provide additional information about the confidence of the player. Then Each of those sub-system can drive independently the avatar' s head.

Regarding the fusions between the audio and video modality, two possibilities are possible: early fusion and late fusion. In the first case, the feature vectors are present simultaneously to a same classifier. In the second case the two feature vectors are presented separately to two different classifiers, and the fusion is done a posteriori. In the case of semantic analysis of video, [11] shows that when late fusion perform better, the difference is more important. Furthermore early fusion has two other disadvantages. The increase of the input feature vector lead to a more complex classifier in terms of number of hidden unit (and so the number of training samples). The constraints of the ReV-TV project involve a low resource method. The second disadvantage is that it is necessary to synchronize the data (the feature vectors) coming from the audio and the video. The late fusion is therefore the chosen solution.

Another way to improve the recognition is to take into account the game context. The use of a priori knowledge, such as game' s situation (answer expectation, good answer, ...) is quite well adapted to the bayesian inference scheme. Figure 1 show an example of decomposition of the French game "Motus" .

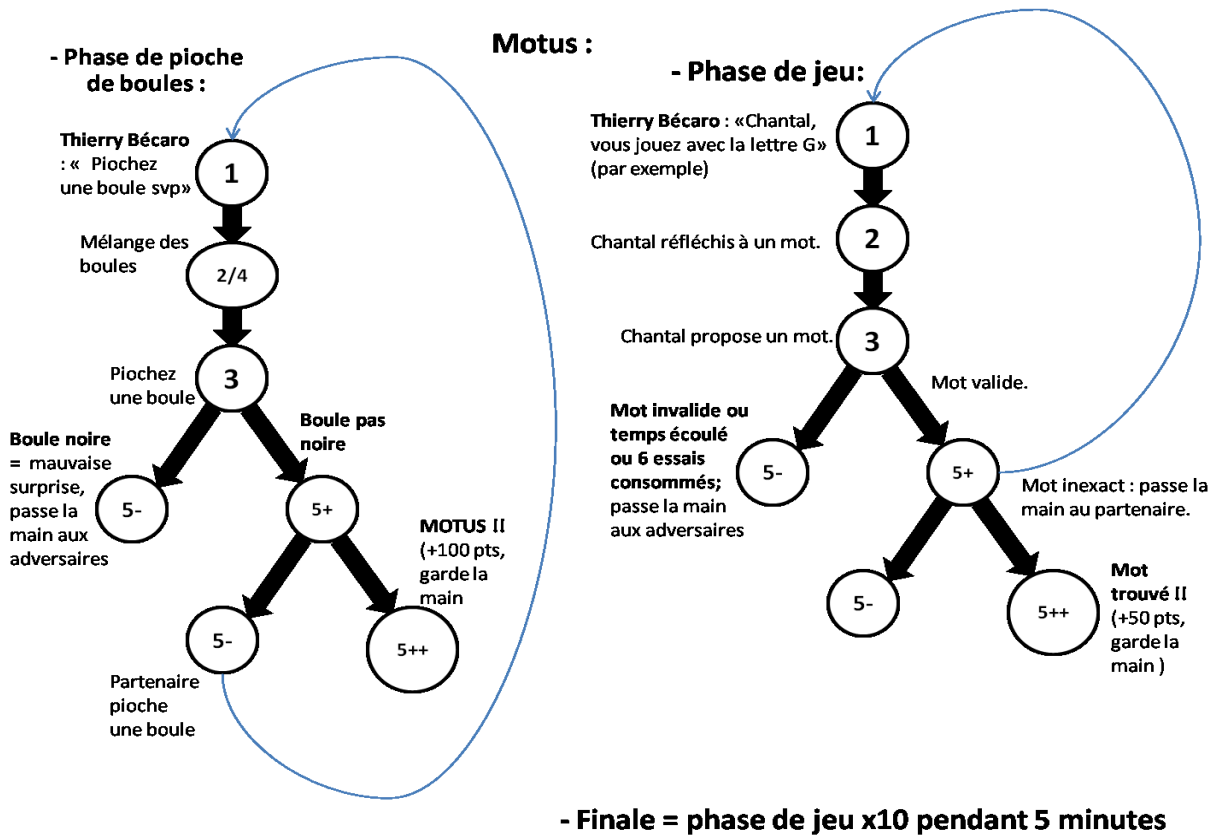


Figure 1: Game dissociated in state for helping emotion classification

5.2 Enhancement of the audio databases

The following axes of enhancement could constitute a base for future work:

- Selection of features among the actual set to provide a representation easier to use
- Complete the databases with other kind of TV-show
- Use of a larger choice of facial feature for the smile and eyebrows :

If smile:

- Wide smile (laugh) (S1)
- Bright smile (S2)
- Timid smile (S3)

If eyebrows:

- Up and fixed eyebrows (surprise)(SL1)
- Frowned and up eyebrows (sadness, deception)(SL2)
- Timid smile (questioning, doubt)(S3)

- As a result of the use of such set of facial feature, a priori knowledge could be added to the animation such as:
 - If S1 detected then SL1 or neutral
 - If S2 detected then SL1 or SL2 or neutral
 - If S3 detected then SL1 or SL2 or SL3 or neutral

- The final system having to process in real-time, the animation quality has to be evaluated subjectively. For this purpose, a mapping to discrete emotions of the avatar of Artefacto has to be found :
 - Smile S2 to joy
 - Eyebrows SL1 to surprise
 - Eyebrow SL2 to sadness
 - Eyebrows SL3 to fear

Bibliography

- [1] C.M. Lee and S.S. Narayaman, "Toward detecting emotions in spoken dialogs", IEEE Trans. On Speech and Audio Processing, vol. 13, no. 2, pp. 293-303, 2005.
- [2] N. F. Fragopanagos and J.G. Taylors, "Emotion recognition in human computer interaction" Neural Networks, vol 18., no. 4, pp. 389 -405,2005.
- [3] P. Ekman, "Facial expression and emotion", 44, 1993, American Psychologist, pp. 352-360.
- [4] R.Cowie and R. Cornelieus, "Describing the emotional states that are expressed in speech", Speech Communication, vol 40, pp. 5-35,2003.
- [5] R. Kehrein, "The prosody of authentic emotions", in Proc. Speech conf., Aix en Provence, FranceApril 2002, pp. 423-426.
- [6] D. Wu, T. D. Parson, S S. Narayanan, "Acoustic Feature Analysis in Speech Emotion Primitives Estimation, InterSpeech, Makuhari, Japan, September, 2010.
- [7] K. P. Truong and D. A. van Leeuwen, "Automatic Detection of Laugther", Interspeech, Lisbon, Portugal, 4-8 September, 2005.
- [8] S. Petridis and M. Pantic, "Audiovisual Laughter Detection Based in Temporal Features", ICMI'08, October 20-22, Chania, Crete, Greece, 2008.
- [9] L. Breiman, "Bagging predictors", Machine Learning, vol. 24 (2), pp. 123-140, 1996.
- [10] J. R. Quinlan, "C4.5: Programs for Machine Learning",. Morgan Kaufmann Publishers, 1993.
- [11] Cees G. M. Snoek , Marcel Worring , Arnold W. M. Smeulders, "Early versus late fusion in semantic video analysis" , Proceedings of the 13th annual ACM international conference on Multimedia, November 06-11, 2005, Hilton, Singapore



**RESEARCH CENTRE
BRETAGNE ATALANTIQUE**

**Campus universitaire de Beaulieu
35042 Rennes Cedex France**

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr
ISSN 0249-6399